Domain Adaptation with Conditional Transferable Components

Mingming Gong[1] MINGMING.GONG@STUDENT.UTS.EDU.AU
Kun Zhang[2,3] KUNZ1@CMU.EDU
Tongliang Liu[1] TLIANG.LIU@GMAIL.COM
Dacheng Tao[1] DACHENG.TAO@UTS.EDU.AU
Clark Glymour[2] CG09@ANDREW.CMU.EDU
Bernhard Schölkopf[3] BS@TUEBINGEN.MPG.DE

[1] Centre for Quantum Computation and Intelligent Systems, FEIT, University of Technology Sydney, NSW, Australia

[2] Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA

[3] Max Plank Institute for Intelligent Systems, Tübingen 72076, Germany

Abstract

Domain adaptation arises in supervised learning when the training (source domain) and test (target domain) data have different distributions. Let X and Y denote the features and target, respectively, previous work on domain adaptation mainly considers the covariate shift situation where the distribution of the features $P(X)$ changes across domains while the conditional distribution $P(Y|X)$ stays the same. To reduce domain discrepancy, recent methods try to find invariant components $\mathcal{T}(X)$ that have similar $P(\mathcal{T}(X))$ on different domains by explicitly minimizing a distribution discrepancy measure. However, it is not clear if $P(Y|\mathcal{T}(X))$ in different domains is also similar when $P(Y|X)$ changes. Furthermore, transferable components do not necessarily have to be invariant. If the change in some components is identifiable, we can make use of such components for prediction in the target domain. In this paper, we focus on the case where $P(X|Y)$ and $P(Y)$ both change in a causal system in which Y is the cause for X. Under appropriate assumptions, we aim to extract conditional transferable components whose conditional distribution $P(\mathcal{T}(X)|Y)$ is invariant after proper location-scale (LS) transformations, and identify how $P(Y)$ changes between domains simultaneously. We provide theoretical analysis and empirical evaluation on both synthetic and real-world data to show the effectiveness of our method.

1. Introduction

Standard supervised learning relies on the assumption that both training and test data are drawn from the same distribution. However, this assumption is likely to be violated in practice if the training and test data are sampled under different conditions. Considering the problem of object recognition, images in different datasets are taken with different cameras or in different imaging conditions (e.g., pose and illumination). In the indoor WiFi localization problem, signals collected during different time periods have different distributions, and one may want to adapt a model trained on the signals received from one time period to the signals collected during other time periods. Domain adaptation approaches aim to solve this kind of problems by transferring knowledge between domains (Pan & Yang, 2010; Jiang, 2008).

To perform domain adaptation, certain assumptions must be imposed in how the distribution changes across domains. For instance, many existing domain adaptation methods consider the covariate shift situation where the distributions on two domains only differ in the marginal distribution of the features $P(X)$, while the conditional distribution of the target given the features $P(Y|X)$ does not change. In this case, one can match the feature distribution $P(X)$ on source and target domains by importance reweighting methods if the source domain is richer than the target domain (Shimodaira, 2000; Sugiyama et al., 2008; Huang et al., 2007). The weights are defined as the density ratio between the source and target domain features and can be efficiently estimated by various methods such as the kernel mean matching procedure (KMM) (Huang et al., 2007). Theoretical analysis of importance reweighting methods for correcting covariate shift has also been studied in (Cortes et al., 2010; Yu & Szepesvári, 2012).

In addition to instance reweighting methods, several state-

Proceedings of the 33^rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

of-the-art approaches try to reduce the domain shift by finding invariant representations or components across domains (Ben-David et al., 2007; Pan et al., 2011; Luo et al., 2014). These invariant components (IC)-type approaches assume that there exist a transformation $\mathcal{T}$ such that $P^{\mathcal{S}}(\mathcal{T}(X)) \approx P^{\mathcal{T}}(\mathcal{T}(X))$, where $P^{\mathcal{S}}$ denotes the source domain distribution and $P^{\mathcal{T}}$ denotes the target domain distribution. To obtain the shared representation, some methods firstly create intermediate representations by projecting the original feature to a series of subspaces and then concatenate them (Gopalan et al., 2011; Gong et al., 2012). Other methods learn a low dimensional projection by explicitly minimizing the discrepancy between the distributions of projected features on source and target domains (Pan et al., 2011; Long et al., 2014; 2015; Baktashmotlagh et al., 2013; Si et al., 2010; 2011; Muandet et al., 2013). Because there are no labels in the target domain in the unsupervised domain adaptation scenario, $\mathcal{T}$ can not be learned by minimizing the distance between $P^{\mathcal{S}}(Y|\mathcal{T}(X))$ and $P^{\mathcal{T}}(Y|\mathcal{T}(X))$. Therefore, these methods simply assume that the transformation $\mathcal{T}$ learned by matching the distribution of transformed features satisfies $P^{\mathcal{S}}(Y|\mathcal{T}(X)) \approx P^{\mathcal{T}}(Y|\mathcal{T}(X))$. However, it is not clear why and when this assumption holds in practice, i.e., under what conditions would $P^{\mathcal{S}}(\mathcal{T}(X)) \approx P^{\mathcal{T}}(\mathcal{T}(X))$ imply $P^{\mathcal{S}}(Y|\mathcal{T}(X)) \approx P^{\mathcal{T}}(Y|\mathcal{T}(X))$? Moreover, the components that are transferable between domains are not necessarily invariant. If the changes in some components are identifiable from the empirical joint distribution on the source domain and the empirical marginal distribution of the features on the target domain, we can make use of these components for domain adaptation.

In fact, to successfully transfer knowledge between domains, one need to capture the underlying causal mechanism, or the data generating process. In particular, for domain adaptation, one would be interested in what types of information are invariant, what types of information change, and how they change across domains. To this end, some recent work address the domain adaptation problem using causal models to characterize how the distribution changes between domains (Schölkopf et al., 2012; Kun Zhang et al., 2013; 2015; Mateo et al., 2016). Let C and E denote the cause and effect, respectively, $P(C)$ characterizes the process which generates the cause and $P(E|C)$ describes the mechanism transforming cause C to effect E. An important feature of a causal system $C \rightarrow E$ is that the mechanism $P(E|C)$ is independent of the cause generating process $P(C)$ (Schölkopf et al., 2012; Janzing & Schölkopf, 2010). For example, in a causal system $X \rightarrow Y$, if $P(Y|X)$ changes across domains, one can hardly correct $P(Y|X)$ unless it is changed by specific transformations like randomly flipping labels (Liu & Tao, 2016), because $P(X)$ contains no information about $P(Y|X)$.

Table 1: Notation used in this paper.

random variable	X	Y
domain	$\mathcal{X}$	$\mathcal{Y}$
observation	x	y
RKHS	$\mathcal{F}$	$\mathcal{G}$
feature map	$\psi(x)$	$\phi(y)$
kernel	$k(x, x')$	$l(y, y')$
kernel matrix on source domain	K	L
source domain data matrix	$\mathbf{x}^{\mathcal{S}}$	$\mathbf{y}^{\mathcal{S}}$
target domain data matrix	$\mathbf{x}^{\mathcal{T}}$	$\mathbf{y}^{\mathcal{T}}$
source domain feature matrix	$\psi(\mathbf{x}^{\mathcal{S}})$	$\phi(\mathbf{y}^{\mathcal{S}})$
target domain feature matrix	$\psi(\mathbf{x}^{\mathcal{T}})$	$\phi(\mathbf{y}^{\mathcal{T}})$

In this paper, we aim to find conditional invariant or transferable components in the generalized target shift (GeTarS) (Kun Zhang et al., 2013) scenario where the causal direction is $Y \rightarrow X$. In this scenario, $P(Y)$ and $P(X|Y)$ change independently to each other, whereas $P(X)$ and $P(Y|X)$ usually change dependently; thus it is possible to correct $P(Y|X)$ from labeled source domain data and unlabeled target domain data. The GeTarS method (Kun Zhang et al., 2013) assumes that all the features can be transferred to the target domain by location-scale (LS) transformation. However, many of the features can be highly noisy or cannot be well matched after LS transformation, which makes GeTarS restrictive in practice. In this paper, under appropriate assumptions, we aim to find the components whose conditional distribution is invariant across domains, i.e., $P^{\mathcal{S}}(\mathcal{T}(X)|Y) \approx P^{\mathcal{T}}(\mathcal{T}(X)|Y)$, and estimate the target domain label distribution $P^{\mathcal{T}}(Y)$ from the labeled source domain and unlabeled target domain. In this way, we can correct the shift in $P(Y|X)$ by using the conditional invariant components and reweighting the source domain data. Similarly, we are able to find the transferable components whose conditional distribution is invariant after proper LS transformations. In addition, we provide theoretical analysis of our method, making clear the assumptions under which the proposed method as well as the previous IC-type methods can work. Finally, we present a computationally efficient method to estimate the involved parameters based on kernel mean embedding of distributions (Smola et al., 2007; Gretton et al., 2012).

2. Conditional Transferable Components

We define conditional invariant components (CIC) X^{ci} as those components satisfying the condition that $P(X^{ci}|Y)$ stays invariant across different domains. Since the location-scale (LS) transformation often occurs in the conditional distribution of the features given the label, we also present the conditional transferable components (CTC) method,

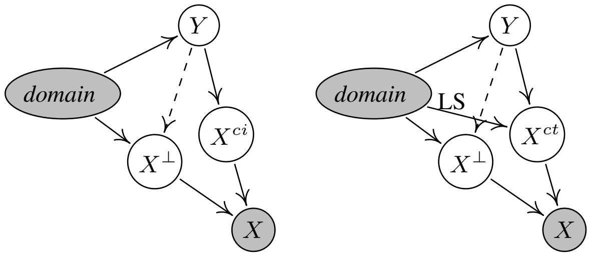


Figure 1: (a) Graphical representation of CIC. Here *domain* denotes the domain-specific selection variable. X^{ci} denotes the components of X whose conditional distribution, $P(X^{ci}|Y)$, is domain-invariant. We assume that X^{ci} can be recovered from X as $\mathcal{T}(X)$. $X^{\perp}$ denotes the remaining components of X; it might be dependent on Y given the domain, and when estimating X^{ci}, we would like such dependence to be as weak as possible so that X^{ci} contains as much information about Y as possible. (b) CTC, where $P(X^{ct}|Y)$ differs only in the location and scale across different domains for each value of Y.

where for each Y value, the conditional distribution of the extracted conditional transferable components X^{ct} given Y, $P(X^{ct}|Y)$, differs only in the location and scale across all domains. Figure 1 gives a simple illustration of the CIC and CTC.

2.1. Conditional Invariant Components

We first assume that there exist d-dimensional conditional invariant components that can be represented as a linear transformation of the D-dimensional raw features, that is, $X^{ci} = W^\intercal X$, where $W \in \mathbb{R}^{D\times d}$ and $X \in \mathbb{R}^D$. To guarantee that there is no redundant information across dimensions of X^{ci}, we constrain the columns of W to be orthonormal:

$$W^\intercal W = I_d. \qquad (1)$$

If we have two domains on which both X and Y are known, we can directly enforce the condition

$$P^{\mathcal{T}}(X^{ci}|Y) = P^{\mathcal{S}}(X^{ci}|Y). \qquad (2)$$

However, in unsupervised domain adaptation, we do not have access to the Y values on the target domain, and thus can not match the conditional distributions directly. Only the empirical marginal distribution of X is available on the test domain.

We will show that under mild conditions, matching the conditional distributions, (2), can be achieved by matching the marginal distribution $P^{\mathcal{T}}(X^{ci})$, which equals to $\int P^{\mathcal{T}}(X^{ci}|y)P^{\mathcal{T}}(y)dy$, with the constructed marginal of X corresponding to $P^{\mathcal{S}}(X^{ci}|Y)$ and $P^{new}(Y)$:

$$P^{new}(X^{ci}) = \int P^{\mathcal{S}}(X^{ci}|y)P^{new}(y)dy. \qquad (3)$$

Definition 1. *A transformation $\mathcal{T}(X)$ is called* `trivial` *if $P\big(\mathcal{T}(X)|Y = v_c\big)$, $c = 1, ..., C$, are linearly dependent.*

With a trivial transformation, the transformed components, $\mathcal{T}(X)$, lose some power for predicting the target Y. For instance, consider a classification problem with only two classes. With a trivial transformation, $P\big(\mathcal{T}(X)|Y = v_c\big)$, $c = 1, 2$, are the same, and as a consequence, $\mathcal{T}(X)$ is not useful for classification.

Fortunately, according to Theorem 1, under mild conditions, if $P^{new}(X^{ci})$ is identical to $P^{\mathcal{T}}(X^{ci})$, i.e.,

$$P^{new}(X^{ci}) = P^{\mathcal{T}}(X^{ci}), \qquad (4)$$

the conditional invariance property of X^{ci}, i.e., condition (2), holds; moreover, the Y distribution on the target domain can also be recovered.

Theorem 1. *Assume that the linear transformation W is* `non-trivial`. *Further assume*

$\mathbf{A}^{\mathbf{CIC}}$**:** *The elements in the set $\big\{\kappa_{c1}P^{\mathcal{S}}(W^\intercal X|Y = v_c) + \kappa_{c2}P^{\mathcal{T}}(W^\intercal X|Y = v_c)\,;\ c = 1, ..., C\big\}$ are linearly independent $\forall\ \kappa_{c1},\ \kappa_{c2}$ ($\kappa_{c1}^2 + \kappa_{c2}^2 \neq 0$), if they are not zero.*

If Eq. 4 holds, we have $P^{\mathcal{S}}(X^{ci}|Y) = P^{\mathcal{T}}(X^{ci}|Y)$ and $p^{new}(Y) = p^{\mathcal{T}}(Y)$, i.e., X^{ci} are conditional invariant components from the source to the target domain.

A complete proof of Theorem 1 can be found in Section S1 of the Supplementary Material. $\mathbf{A}^{\mathbf{CIC}}$ is enforced to ensure that the changes in the weighted conditional distributions $P(X^{ci}|Y = v_c)P(Y = v_c),\ c = 1, ..., C$ are linearly independent, which is necessary for correcting joint distributions by matching marginal distributions of features. Theorem 1 assumes that the distributions on different domains can be perfectly matched. However, it is difficult to find such ideal invariant components in practice. In Section 3, we will show that the distance between the joint distributions across domains can be bounded by the distance between marginal distributions of features across domains under appropriate assumptions.

Now we aim to find a convenient method to enforce the condition (4). Assume that $P^{\text{new}}(Y)$ is absolutely continuous w.r.t. $P^{\mathcal{S}}(Y)$. We can represent $P^{\text{new}}(Y = y)$ as $P^{\mathcal{T}}(y) = \beta(y)P^{\mathcal{S}}(y)$, where $\beta(y)$ is a density ratio, satisfying $\beta(y) \geq 0$ and $\int \beta(y)P^{\mathcal{S}}(y)dy = 1$, since both $P^{\text{new}}(Y)$ and $P^{\mathcal{S}}(Y)$ are valid distribution density or mass functions. Let $\beta_i \triangleq \beta(y_i^{\mathcal{S}})$, and $\boldsymbol{\beta} = [\beta_1, .., \beta_{n^{\mathcal{S}}}]^\intercal \in \mathbb{R}^{n^{\mathcal{S}}}$; they satisfy the constraint

$$\beta_i \geq 0, \text{ and } \sum_{i=1}^{n^{\mathcal{S}}} \beta_i = n^{\mathcal{S}}. \qquad (5)$$

A method to achieve (4) is to minimize the squared maximum mean discrepancy (MMD):

$$\begin{aligned}&\Big|\Big|\mu_{P^{new}(X^{ci})}[\psi(X^{ci})] - \mu_{P^{\mathcal{T}}(X^{ci})}[\psi(X^{ci})]\Big|\Big|^2 \qquad (6)\\ =&\Big|\Big|\mathbb{E}_{X^{ci}\sim P^{new}(X^{ci})}[\psi(X^{ci})] - \mathbb{E}_{X^{ci}\sim P^{\mathcal{T}}(X^{ci})}[\psi(X^{ci})]\Big|\Big|^2.\end{aligned}$$

One way to enforce this condition is to exploit the embedding of the conditional distribution $P^{\mathcal{S}}(X^{ci}|Y)$ as the bridge to connect the two involved quantities, as in (Kun Zhang et al., 2013). However, we will show it is possible to develop a simpler procedure[1].

Because of (3), we have

$$\begin{aligned}&\mathbb{E}_{X^{ci}\sim P^{new}(X^{ci})}[\psi(X^{ci})] = \int \psi(x^{ci})P^{new}(x^{ci})dx^{ci}\\ =&\int \psi(x^{ci})P^{\mathcal{S}}(x^{ci}|y)P^{\mathcal{S}}(y)\beta(y)dydx^{ci}\\ =&\int \psi(x^{ci})\beta(y)P^{\mathcal{S}}(y,x^{ci})dydx^{ci}\\ =&\mathbb{E}_{(Y,X^{ci})\sim P^{\mathcal{S}}(Y,X^{ci})}[\beta(Y)\psi(X^{ci})]. \qquad (7)\end{aligned}$$

As a consequence, (6) reduces to

$$J^{ci} = \Big|\Big|\mathbb{E}_{(Y,X)\sim p^{\mathcal{S}}}[\beta(Y)\psi(W^{\intercal}X)] - \mathbb{E}_{X\sim p^{\mathcal{T}}}[\psi(W^{\intercal}X)]\Big|\Big|^2.$$

In practice, we minimize its empirical version

$$\begin{aligned}\hat{J}^{ci} =&\Big|\Big|\frac{1}{n^{\mathcal{S}}}\psi(W^{\intercal}\mathbf{x}^{\mathcal{S}})\boldsymbol{\beta} - \frac{1}{n^{\mathcal{T}}}\psi(W^{\intercal}\mathbf{x}^{\mathcal{T}})\mathbf{1}\Big|\Big|^2\\ =&\frac{1}{{n^{\mathcal{S}}}^2}\boldsymbol{\beta}^{\intercal}K_W^{\mathcal{S}}\boldsymbol{\beta} - \frac{2}{n^{\mathcal{S}}n^{\mathcal{T}}}\mathbf{1}^{\intercal}K_W^{\mathcal{T},\mathcal{S}}\boldsymbol{\beta} + \frac{1}{{n^{\mathcal{T}}}^2}\mathbf{1}^{\intercal}K_W^{\mathcal{T}}\mathbf{1}\end{aligned}$$

where $\boldsymbol{\beta} \triangleq \beta(\mathbf{y}^{\mathcal{S}})$, $\mathbf{1}$ is the $n^{\mathcal{T}} \times 1$ vector of ones, $K_W^{\mathcal{T}}$ and $K_W^{\mathcal{S}}$ denote the kernel matrix on $W^{\intercal}\mathbf{x}^{\mathcal{T}}$ and $W^{\intercal}\mathbf{x}^{\mathcal{S}}$, respectively, and $K_W^{\mathcal{T},\mathcal{S}}$ the cross kernel matrix between $W^{\intercal}\mathbf{x}^{\mathcal{T}}$ and $W^{\intercal}\mathbf{x}^{\mathcal{S}}$. Note that the kernel has to be characteristic; otherwise there are always trivial solutions. In this paper, we adopt the Gaussian kernel function, which has been shown to be characteristic (Sriperumbudur et al., 2011).

2.2. Location-Scale Conditional Transferable Components

In practice, one may not find sufficient conditional invariant components which also have high discriminative power. To discover more useful conditional transferable components, we assume here that there exist transferable components that can be approximated by a location-scale transformation across domains. More formally, we assume that there exist W, $\mathbf{a}(Y^{\mathcal{S}}) = [a_1(Y^{\mathcal{S}}),...,a_d(Y^{\mathcal{S}})]^{\intercal}$ and $\mathbf{b}(Y^{\mathcal{S}}) = [b_1(Y^{\mathcal{S}}),...,b_d(Y^{\mathcal{S}})]^{\intercal}$, such that the conditional distribution of $X^{ct} \triangleq \mathbf{a}(Y^{\mathcal{S}}) \circ (W^{\intercal}X^{\mathcal{S}}) + \mathbf{b}(Y^{\mathcal{S}})$ given $Y^{\mathcal{S}}$ is close to that of $W^{\intercal}X^{\mathcal{T}}$ given $Y^{\mathcal{T}}$. The transformed training data matrix $\mathbf{x}^{ct} \in \mathbb{R}^{d\times n^{\mathcal{S}}}$ can be written in matrix form

$$\mathbf{x}^{ct} = \mathbf{A} \circ \left(W^{\intercal}\mathbf{x}^{\mathcal{S}}\right) + \mathbf{B}, \qquad (8)$$

where $\circ$ denotes the Hadamard product, the i-th columns of $\mathbf{A} \in \mathbb{R}^{d\times n^{\mathcal{S}}}$ and $\mathbf{B} \in \mathbb{R}^{d\times n^{\mathcal{S}}}$ are $\mathbf{a}(y_i)$ and $\mathbf{b}(y_i)$, respectively. Using (8), we can generalize J^{ci} to

$$J^{ct} =\Big|\Big|\mathbb{E}_{(Y,X^{ct})\sim p^{\mathcal{S}}}[\beta(Y)X^{ct}] - \mathbb{E}_{X\sim p^{\mathcal{T}}}[\psi(W^{\intercal}X)]\Big|\Big|^2,$$

and its empirical version $\hat{J}^{ci}$ to

$$\begin{aligned}\hat{J}^{ct} =&\Big|\Big|\frac{1}{n^{\mathcal{S}}}\psi\left(\mathbf{x}^{ct}\right)\boldsymbol{\beta} - \frac{1}{n^{\mathcal{T}}}\psi(W^{\intercal}\mathbf{x}^{\mathcal{T}})\mathbf{1}\Big|\Big|^2\\ =&\frac{1}{{n^{\mathcal{S}}}^2}\boldsymbol{\beta}^{\intercal}\tilde{K}^{\mathcal{S}}\boldsymbol{\beta} - \frac{2}{n^{\mathcal{S}}n^{\mathcal{T}}}\mathbf{1}^{\intercal}\tilde{K}^{\mathcal{T},\mathcal{S}}\boldsymbol{\beta} + \frac{1}{{n^{\mathcal{T}}}^2}\mathbf{1}^{\intercal}K_W^{\mathcal{T}}\mathbf{1}\end{aligned}$$

where $\tilde{K}^{\mathcal{S}}$ denote the kernel matrix on $\mathbf{x}^{ct}$ and $\tilde{K}^{\mathcal{T},\mathcal{S}}$ the cross kernel matrix between $W^{\intercal}\mathbf{x}^{\mathcal{T}}$ and $\mathbf{x}^{ct}$. The identifiability of $\mathbf{A}$ and $\mathbf{B}$ can be easily obtained by combing the results of Theorem 1 in this paper and Theorem 2 in (Kun Zhang et al., 2013). In practice, we expect the changes in the conditional distribution of X^{ct} given $Y^{\mathcal{S}}$ to be as small as possible. Thus we add a regularization term on $\mathbf{A}$ and $\mathbf{B}$, i.e., $J^{reg} = \frac{\lambda_S}{n^{\mathcal{S}}}||\mathbf{A} - \mathbf{1}_{d\times n^{\mathcal{S}}}||_F^2 + \frac{\lambda_L}{n^{\mathcal{S}}}||\mathbf{B}||_F^2$, where $\mathbf{1}_{d\times n^{\mathcal{S}}}$ is the $d \times n^{\mathcal{S}}$ matrix of ones.

2.3. Target Information Preservation

At the same time, because the components X^{ct} will be used to predict Y, we would like X^{ct} to preserve the information about Y. The information in the given feature X about the Y is completely preserved in the components X^{ct} if and only if $Y \perp\!\!\!\perp X \,|\, X^{ct}$. We adopt the kernel dimensionality reduction framework (Fukumizu et al., 2004) to achieve so. It has been shown that $Y \perp\!\!\!\perp X \,|\, X^{ct} \iff \Sigma_{YY|X^{ct}} - \Sigma_{YY|X} = 0$, where $\Sigma_{YY|X}$ is the conditional covariance operator on $\mathcal{G}$.

Consequently, to minimize the conditional dependence between Y and X given X^{ct}, one can minimize the determinant of trace of $\Sigma_{YY|X^{ct}}$. Here we use a slightly simpler estimator for its trace. According to its definition (Baker, 1973), $\Sigma_{YY|X^{ct}} = \Sigma_{YY} - \Sigma_{Y,X^{ct}}\Sigma^{-1}_{X^{ct},X^{ct}}\Sigma_{X^{ct},Y}$, where $\Sigma_{\cdot\cdot}$ is the covariance or cross-covariance operator.

We can use $\frac{1}{n^{\mathcal{S}}}\phi(\mathbf{y}^{\mathcal{S}})\phi^{\intercal}(\mathbf{y}^{\mathcal{S}})$, $\frac{1}{n^{\mathcal{S}}}\phi(\mathbf{y}^{\mathcal{S}})\psi^{\intercal}(\mathbf{x}^{ct})$, and $\frac{1}{n^{\mathcal{S}}}\psi(\mathbf{x}^{ct})\psi^{\intercal}(\mathbf{x}^{ct})$ as the estimators of Σ_{YY}, $\Sigma_{Y,X^{ct}}$, and $\Sigma_{X^{ct},X^{ct}}$, respectively, on the source-domain data. Consequently, on such data we have the estimator

$$\mathbb{T}\mathrm{r}[\hat{\Sigma}_{YY|X^{ct}}]$$

[1] Alternatively, one may make use of the kernel mean embedding of conditional distributions in the derivation, as in the algorithm for correcting target shift (Kun Zhang et al., 2013), but it will be more complex. Likewise, by making use of (7), the objective function used there can be simplified: in their equation (5), the matrix $\mathbf{\Omega}$ can be dropped.

$$
\begin{aligned}
=&\mathbb{T}\mathrm{r}[\hat{\Sigma}_{YY}] - \mathbb{T}\mathrm{r}[\hat{\Sigma}_{Y,X^{ct}}\hat{\Sigma}^{-1}_{X^{ct},X^{ct}}\hat{\Sigma}_{X^{ct},Y}] \\
=&\frac{1}{n^{\mathcal{S}}}\mathbb{T}\mathrm{r}[\phi(\mathbf{y}^{\mathcal{S}})\phi^{\intercal}(\mathbf{y}^{\mathcal{S}})] - \frac{1}{n^{\mathcal{S}}}\mathbb{T}\mathrm{r}[\phi(\mathbf{y}^{\mathcal{S}})\psi^{\intercal}(\mathbf{x}^{ct})\cdot \\
&\big(\psi(\mathbf{x}^{ct})\psi^{\intercal}(\mathbf{x}^{ct}) + n^{\mathcal{S}}\varepsilon I\big)^{-1}\cdot\psi(\mathbf{x}^{ct})\phi^{\intercal}(\mathbf{y}^{\mathcal{S}})] \\
=&\varepsilon\mathbb{T}\mathrm{r}[L(\tilde{K}^{\mathcal{S}} + n^{\mathcal{S}}\varepsilon I)^{-1}], \qquad (9)
\end{aligned}
$$

where ε is a regularization parameter to prevent ill conditions on the matrix inverse and is set to 0.01 in our experiments.

2.4. Reparameterization of $\boldsymbol{\beta}$, A, and B

By combining $\hat{J}^{ct}$, J^{reg}, and $\mathbb{T}\mathrm{r}[\hat{\Sigma}_{YY|X^{ct}}]$, we aim to estimate the parameters $\boldsymbol{\beta}$, W, $\mathbf{A}$, and $\mathbf{B}$ by minimizing

$$
\hat{J}^{ct}_{con} = \hat{J}^{ct} + \lambda\mathbb{T}\mathrm{r}[\hat{\Sigma}_{YY|X^{ct}}] + J^{reg} \qquad (10)
$$

under constraints (1) and (5). However, we cannot directly minimize (10) with respect to $\boldsymbol{\beta}$, $\mathbf{A}$, and $\mathbf{B}$ because β, $\mathbf{a}$, and $\mathbf{b}$ are functions of y. Thus, we reparametrize $\boldsymbol{\beta}$, $\mathbf{A}$, and $\mathbf{B}$ with new parameters. In this paper, we focus on the case where Y is discrete. Let C be the cardinality of $\mathcal{Y}$ and denote by $v_1, ..., v_C$ its possible values. Let n_c denotes number of examples with $Y = v_c$, we can define a matrix $R^{dis} \in \mathbb{R}^{n^{\mathcal{S}}\times C}$ where R^{dis}_{ic} is $\frac{n_{\mathcal{S}}}{n_c}$ if $y_i = v_c$ and is zero everywhere else. $\boldsymbol{\beta}$ can then be reparameterized as $\boldsymbol{\beta} = R^{dis}\boldsymbol{\alpha}$, where the $\boldsymbol{\alpha} \in \mathbb{R}^{C\times 1}$ is the new parameter, providing a compact representation for $\boldsymbol{\beta}$. Similarly, $\mathbf{A}$ and $\mathbf{b}$ can be reparameterized as $(R^{dis}\mathbf{G})^{\intercal}$ and $(R^{dis}\mathbf{H})^{\intercal}$, where $\mathbf{G} \in \mathbb{R}^{C\times d}$ and $\mathbf{H} \in \mathbb{R}^{C\times d}$ are the effective parameters. The constraint on $\boldsymbol{\beta}$, (5), is equivalent to the corresponding constraint on $\boldsymbol{\alpha}$:

$$
[R^{dis}\boldsymbol{\alpha}]_i \geq 0, \text{ and } \mathbf{1}^{\intercal}\boldsymbol{\alpha} = 1. \qquad (11)
$$

2.5. Optimization

We estimate the parameters $\boldsymbol{\alpha}$, W, $\mathbf{G}$, and $\mathbf{H}$ by minimizing $\hat{J}^{ct}_{con}$ under constraints (1) and (11). We iteratively alternate between minimizing $\boldsymbol{\alpha}$, W, and $[\mathbf{G}, \mathbf{H}]$. For the CIC method, we only optimize W and $\boldsymbol{\alpha}$ by fixing $\mathbf{G}$ and $\mathbf{H}$. For $\boldsymbol{\alpha}$, we use quadratic programming (QP) to minimize $\hat{J}^{ct}_{con}$ w.r.t. $\boldsymbol{\alpha}$ under constraint (11). When minimizing $\hat{J}^{ct}_{con}$ w.r.t. W, one should guarantee that W is on the Grassmann manifold, as implied by constraint (1). Therefore, we find W by the conjugate gradient algorithm on the Grassmann manifold , which is an efficient approach by exploiting the geometric properties of orthogonality and rotation invariance (Edelman et al., 1999). $[\mathbf{G}, \mathbf{H}]$ can be found by standard conjugate gradient optimization procedure. The derivation of the required derivatives is given in the Section S5 of the Supplementary Materials.

3. Theoretical Analysis

We theoretically analyze our CIC method by developing a bound relating source and target domain expected errors. The analysis of the CTC method can be performed in a similar way. Current analysis methods on domain adaptation (Ben-David et al., 2007; 2010) decompose the joint distribution $P(X, Y)$ to $P(X)P(Y|X)$ and measure their distance between domains separately. Therefore, many existing methods explicitly minimizes the discrepancy between source and target domains by learning invariant components $X^{ci} = W^{\intercal}X$ with similar marginal distributions $p^{\mathcal{S}}(X^{ci}) \approx p^{\mathcal{T}}(X^{ci})$. However, it is not sure whether the distance between $P^{\mathcal{S}}(Y|X^{ci})$ and $P^{\mathcal{T}}(Y|X^{ci})$ is also small.

We will show that, in the $Y \rightarrow X$ situation, the distance between the joint distributions across domains can be bounded by the distance between marginal distributions of features across domains , if the assumptions in Theorem 1 holds. Different from previous works, we decompose the joint distribution in the causal direction, i.e., $P(X^{ci}, Y) = P(X^{ci}|Y)P(Y)$. Following (Ben-David et al., 2007; 2010), we only consider the binary classification problem with $1 - 0$ loss for convenience.

Before stating the main theorems, we first introduce the following Lemma. It is similar to Theorem 1 in (Ben-David et al., 2010), but we directly measure the distance between joint distributions on different domains instead of separately measuring the distance between $P^{\mathcal{S}}(X)$ and $P^{\mathcal{T}}(X)$ and the distance between $P^{\mathcal{S}}(Y|X)$ and $P^{\mathcal{T}}(Y|X)$.

Lemma 1. *For a hypothesis $h \in \mathcal{H}$, let $\epsilon_{new}(h)$ and $\epsilon_{\mathcal{T}}(h)$ be the expected error w.r.t. 1-0 loss on the constructed new domain and target domain respectively. We have*

$$
\epsilon_{\mathcal{T}}(h) \leq \epsilon_{new}(h) + d_1(p^{new}(X^{ci}, Y), p^{\mathcal{T}}(X^{ci}, Y)), \quad (12)
$$

where $d_1(p^{new}(X^{ci}, Y), p^{\mathcal{T}}(X^{ci}, Y))$ is the L^1 or variation divergence defined in (Ben-David et al., 2010).

The proof of Lemma 1 is given in Section S2 of the Supplementary Material.

Because d_1 is difficult to calculate in practice, we measure distribution discrepancy between the joint distribution on the new domain and the target domain by squared MMD distance, i.e.,

$$
\begin{aligned}
&d_k(p^{\text{new}}(X^{ci}, Y), p^{\mathcal{T}}(X^{ci}, Y)) \\
=&\big|\big|\mathbb{E}_{(X^{ci},Y)\sim P^{\text{new}}(X^{ci},Y)}[\psi(X^{ci})\otimes\phi(Y)] \\
&- \mathbb{E}_{(X^{ci},Y)\sim P^{\mathcal{T}}(X^{ci},Y)}[\psi(X^{ci})\otimes\phi(Y)]\big|\big|^2, \qquad (13)
\end{aligned}
$$

where $\otimes$ denotes the tensor product.

The following theorem states that the distance between the source and target domain joint distribution can be bounded by the distance between the source and target domain marginal distribution of X^{ci} under certain assumptions.

Theorem 2. *Let Δ_c denote*

$$\begin{aligned}\Delta_c =& P^{new}(Y=c)\mu_{P^{\mathcal{S}}(X^{ct}|Y=c)}[\psi(X^{ct})] \\ &- P^{\mathcal{T}}(Y=c)\mu_{P^{\mathcal{T}}(X^{ct}|Y=c)}[\psi(X^{ct})],\ c=0,1,\end{aligned}$$

and θ denote the angle between Δ_0 and Δ_1. If W is `non-trivial` *and $\mathbf{A}^{\mathbf{CIC}}$ holds, i.e., $0<\theta<\pi$,*

$$\begin{aligned}d_k(p^{new}(X^{ci},Y),p^{\mathcal{T}}(X^{ci},Y)) \leq& J^{ci}1_{0<\theta\leq\pi/2} \\ &+\frac{2J^{ci}}{\sin^2\theta}1_{\pi/2<\theta<\pi},\end{aligned}$$

where $1_{\{\cdot\}}$ denotes the indicator function.

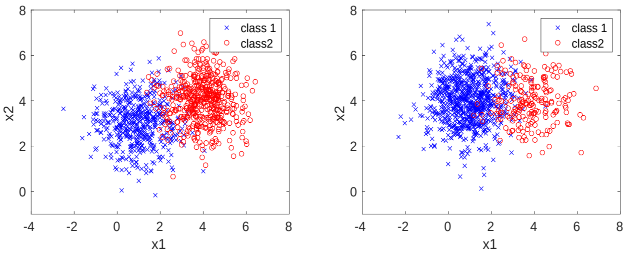
The proof of Theorem 2 can be found in Section S3 of the Supplementary Material.

Remark Suppose we have found the ideal $\boldsymbol{\beta}$ such that $P^{\text{new}}(Y)=P^{\mathcal{T}}(Y)$, then Δ_1 and Δ_0 represent the changes in conditional distribution $P(X^{ci}|Y=1)$ and $P(X^{ci}|Y=0)$, respectively. If one can find perfectly invariant components, i.e., $J^{ci}=0$, which implies $\Delta_1+\Delta_0=0$. If $\mathbf{A}^{\mathbf{CIC}}$ is violated, that is Δ_1 and Δ_0 can be linearly dependent if they are not zeros, then one cannot expect that the conditional distribution $P(X^{ci}|Y)$ is invariant, i.e., $\Delta_1=0$ and $\Delta_0=0$. In this case, the conditional distributions $P(X^{ci}|Y=1)$ and $P(X^{ci}|Y=0)$ change dependently to make the marginal distribution $P(X^{ci})$ invariant across domains. This usually happens in the $X\rightarrow Y$ situation, while rarely happens in the $Y\rightarrow X$ situation. If $\mathbf{A}^{\mathbf{CIC}}$ is violated, it can be seen from Theorem 2 that d_k cannot be bounded by J^{ci} when $\theta=\pi$. Interestingly, when the changes in $P(X^{ci}|Y=1)$ and $P(X^{ci}|Y=0)$ do not cancel each other, i.e., $0<\theta\leq\pi/2$, d_k can be tightly bounded by J^{ci} which can be estimated from labeled data in the source domain and unlabeled data in the target domain.

In practice, we optimize $\hat{J}^{ci}$ w.r.t. W and $\boldsymbol{\alpha}$ under constraints (11) and (1). Let $\boldsymbol{\alpha}_n$ and W_n be the learned parameter according to $\hat{J}^{ci}$. Since the objective function is non-convex w.r.t. W, we cannot expect W_n to converge to the optimal one. However, the optimality of the parameter $\boldsymbol{\alpha}$ can be obtained. We will provide an upper bound for the following defect $J^{ci}(\boldsymbol{\alpha}_n,W_n)-J^{ci}(\boldsymbol{\alpha}^*,W_n)$, where $\boldsymbol{\alpha}^*$ denotes the optimal one.

Theorem 3. *Assume the RKHS employed are bounded such that $||\psi(x)||_2\leq\wedge_2$ for all $x\in\mathcal{X}$. For any $\delta>0$, with probability at least $1-\delta$, we have*

$$\begin{aligned}J^{ci}(\boldsymbol{\alpha}_n,W_n)-J^{ci}(\boldsymbol{\alpha}^*,W_n)\leq 8\wedge_2^2\Bigg(2\sqrt{\sum_{c=1}^{C}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}}} \\ +8\sqrt{\frac{1}{2}\log\frac{2}{\delta}(\max_{c\in\{1,\ldots,C\}}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}})}\Bigg)^{\frac{1}{2}}.\end{aligned}$$



(a) source domain (b) target domain

Figure 2: Toy data to illustrate the difference between DIP and CIC: (a) The source domain data; (b) The target domain data.

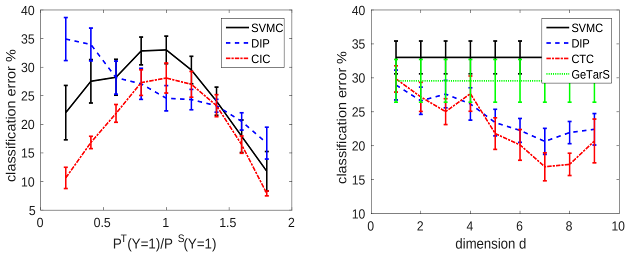
The proof of Theorem 3 can be found in Section S4 of the Supplementary Material.

4. Relation to IC-type Methods

If $P(Y)$ stays the same across domains, the CIC method reduces to one of the IC-type methods: domain invariant projection (DIP) (Baktashmotlagh et al., 2013). However, their motivations are quite different. IC-type methods, which were proposed for correction of covariate shift, aim to find components X^{ci} whose distribution $P(X^{ci})$ is invariant across domains. Since $P(Y|X)$ stays the same in the covariate shift, $p(Y|X^{ci})$ also stays the same. However, if $P(Y|X)$ changes, it is not sure whether $P(Y|X^{ci})$ could stay the same.

We find that IC-type methods can actually be considered as a way to achieve our CIC method under target shift, given that the distribution $P(Y)$ remains the same across domains. According to Theorem 1, if $P^{\mathcal{S}}(Y)=P^{new}(Y)$, we have $P^{\mathcal{S}}(X^{ci},Y)=P^{\mathcal{T}}(X^{ci},Y)$ and thus $P^{\mathcal{S}}(Y|X^{ci})=P^{\mathcal{T}}(Y|X^{ci})$. In other words, under assumption $\mathbf{A}^{\mathbf{CIC}}$, if $P(Y)$ stays the same across domains, $P^{\mathcal{S}}(X^{ci})=P^{\mathcal{T}}(X^{ci})$ leads to $P^{\mathcal{S}}(Y|X^{ci})=P^{\mathcal{T}}(Y|X^{ci})$.

If $P(Y)$ changes, CIC and DIP usually lead to different results. Suppose there exist some components of X, X^{ci}, whose conditional distribution given Y stay the same across domains. In general, when $P(Y)$ changes across domains, it is very unlikely for X^{ci} to have domain-invariant distributions. As illustrated in Figure 2, the conditional distributions $P(X_1|Y=1)$, $P(X_1|Y=2)$, and $P(X_2|Y=2)$ do not change across domains, while the conditional distribution $P(X_2|Y=1)$ is changed by shifting its mean from 3 to 4. The class prior $P(Y=1)$ on the source and target domain is 0.5 and 0.8, respectively. Thus X_1 is a conditional invariant component while X_2 is not. We evaluate the squared MMD between the marginal distribution of these two components. DIP gives the results of $\text{MMD}^2_{X_1}=7.72e-2$ and $\text{MMD}^2_{X_2}=2.38e-2$ and CIC gives $\text{MMD}^2_{X_1}=2.25e-4$ and $\text{MMD}^2_{X_2}=6.44e-2$. That is to say, DIP wrongly considers X_2 as conditional in-



(a) class ratio

(b) dimension d

Figure 3: Performance comparison on simulated data: (a) Classification error w.r.t. class ratio; (b) Classification error w.r.t. dimension d.

variant component, while CIC considers X_1 as conditional invariant component correctly.

5. Experiments

In this section we present experimental results on both simulated and real data to show the effectiveness of the proposed CIC and CTC method. We select the hyperparameters of our methods as follows. For Gaussian kernel used in MMD, we set the standard deviation parameter σ to the median distance between all source examples. The regularization parameters of the LS transformation are set to $\lambda_S = 0.001$ and $\lambda_L = 0.0001$. We choose different parameters for location and scale transformations because we find that the conditional distributions usually have larger location changes. The regularization parameter for the target information preserving (TIP) term is set to $\lambda = 0.001$, resulting in two regularized methods: CIC-TIP and CTC-TIP. We use $\boldsymbol{\beta}$-weighted support vector machine (SVM) and weighted kernel ridge regression (KRR) for classification and regression problems, respectively. For details, please refer to (Kun Zhang et al., 2013). We use linear kernel for simulation data and Gaussian kernel for real data.

5.1. Simulations

We generate binary classification training and test data from a 10-dimensional mixture of Gaussians:

$$x \sim \sum_{i=1}^{2} \pi_i \mathcal{N}(\boldsymbol{\theta}_i, \Sigma_i), \;\; \boldsymbol{\theta}_{ij} \sim \mathcal{U}(-0.25, 0.25)$$
$$\Sigma_i \sim 0.01 * \mathcal{W}(2 \times \mathbf{I}_D, 7), \qquad (14)$$

where $\mathcal{U}(a, b)$ and $\mathcal{W}(\Sigma, df)$ represent the uniform distribution and Wishart distribution, respectively. The cluster indices are used as the ground truth class labels. We apply two types of transformations to the test data to make the test and training data have different distributions. Firstly, we randomly apply LS transformation on 5 randomly selected features for each class. In addition, we apply affine transformation on another 2 randomly chosen features. The remaining 3 features are left unchanged to ensure that the IC-type methods will not fail on the transformed data.

Table 2: Comparison of different methods on the Office+Caltech256 dataset.

	SVM	GFK	TCA	LM	GeTarS	DIP	DIP-CC	CTC	CTC-TIP
A→C	41.7	42.2	35.0	45.5	44.9	47.4	47.2	48.6	**48.8**
A→D	41.4	42.7	36.3	47.1	45.9	50.3	49.0	**52.9**	52.2
A→W	34.2	40.7	27.8	46.1	39.7	47.5	47.8	**49.8**	49.1
C→A	51.8	44.5	41.4	56.7	56.9	55.7	**58.7**	58.1	57.9
C→D	54.1	43.3	45.2	57.3	49.0	60.5	**61.2**	59.2	58.5
C→W	46.8	44.7	32.5	49.5	46.4	58.3	58.0	**58.6**	57.8
W→A	31.1	31.8	24.2	40.2	38.4	42.6	40.9	**43.2**	43.1
W→C	31.5	30.8	22.5	35.4	34.3	34.2	37.2	38.3	**38.8**
W→D	70.7	75.6	80.2	75.2	86.0	88.5	91.7	**94.3**	93.6

We compare our methods against domain invariant projection (DIP) (Baktashmotlagh et al., 2013), which is equivalent to our CIC method when $P(Y)$ does not change. We also include the GeTarS method (Kun Zhang et al., 2013) which assumes that all the features are transferable by LS-transformation. The regularization parameter C of SVM are selected by 5-fold cross validation on a grid. We repeat the experiments for 20 times and report the average classification error.

Firstly, we test the methods' sensitiveness to changes in class prior probability $P(Y)$. we set the source class prior $P^{\mathcal{S}}(Y = 1)$ to 0.5 and the number of components d to 5. The target domain class prior $p^{\mathcal{T}}(Y = 1)$ varies from 0.1 to 0.9 and the corresponding class ratio $\beta_1 = p^{\mathcal{T}}(Y = 1)/P^{\mathcal{S}}(Y = 1)$ is $0.2, 0.4, ..., 1.8$. We compare CIC and DIP which all aim at finding invariant components. Figure 3 (a) gives the classification error as β_1 ranges from 0.2 to 1.8. We can see that the performance of DIP decreases as β_1 gets far away from 1, while CIC performs well with all the β_1 values. We can also see that DIP outperforms CIC when $P(Y)$ changes slightly, which is reasonable because CIC introduces random error in the estimation of $\boldsymbol{\beta}$.

Secondly, we evaluate the effectiveness of discovering transferable components with LS transformation. We set the prior distribution on both domains to $P^{\mathcal{S}}(Y = 1) = p^{\mathcal{T}}(y = 1) = 0.5$ and demonstrate how the performances vary with the dimensionality d of the learned components. Figure 3 (b) shows the classification error of each method as d ranges from 1 to 9. We can see that CTC outperforms DIP when $d > 4$, indicating that CTC successfully matches the features transformed by LS transformation for domain transfer. GeTarS does not perform well because LS transformation fails to match the two affine-transformed features.

5.2. Object Recognition

We also compare our approaches with alternatives on the Office-Caltech dataset introduced in (Gong et al., 2012). The Office-Caltech dataset was constructed by extracting the 10 categories common to the Office dataset (Saenko

Table 3: Comparison of different methods on the WiFi dataset.

	KRR	TCA	SuK	DIP	DIP-CC	GeTarS	CTC	CTC-TIP
t1 → t2	80.84 ± 1.14	86.85 ± 1.1	90.36 ± 1.22	87.98 ± 2.33	$\mathbf{91.30 \pm 3.24}$	86.76 ± 1.91	89.36 ± 1.78	89.22 ± 1.66
t1 → t3	76.44 ± 2.66	80.48 ± 2.73	$\mathbf{94.97 \pm 1.29}$	84.20 ± 4.29	84.32 ± 4.57	90.62 ± 2.25	94.80 ± 0.87	92.60 ± 4.50
t2 → t3	67.12 ± 1.28	72.02 ± 1.32	85.83 ± 1.31	80.58 ± 2.10	81.22 ± 4.31	82.68 ± 3.71	87.92 ± 1.87	$\mathbf{89.52 \pm 1.14}$
hallway1	60.02 ± 2.60	65.93 ± 0.86	76.36 ± 2.44	77.48 ± 2.68	76.24 ± 5.14	84.38 ± 1.98	$\mathbf{86.98 \pm 2.02}$	86.78 ± 2.31
hallway2	49.38 ± 2.30	62.44 ± 1.25	64.69 ± 0.77	78.54 ± 1.66	77.8 ± 2.70	77.38 ± 2.09	87.74 ± 1.89	$\mathbf{87.94 \pm 2.07}$
hallway3	48.42 ± 1.32	59.18 ± 0.56	65.73 ± 1.57	75.10 ± 3.39	73.40 ± 4.06	80.64 ± 1.76	$\mathbf{82.02 \pm 2.34}$	81.72 ± 2.25

et al., 2010) and the Caltech256 (Griffin et al., 2007) dataset. We have four domains in total: Amazon (images downloaded from Amazon), Webcam (low-resolution images by a web camera), DSLR (high-resolution images by a SLR camera), and Caltech-256. We use the bag of visual words features provided by (Gong et al., 2013) for our evaluation.

In our experiments, we use the evaluation protocol in (Gong et al., 2013). We compare CTC and CTC-TIP with several state-of-the-art methods: geodesic flow kernel (GFK) (Gong et al., 2012), transfer component analysis (TCA) (Pan et al., 2011), landmark selection (LM) (Gong et al., 2013), DIP and its cluster regularized version DIP-CC, and GeTarS. The dimensionality of the of the projection matrix W is determined by the subspace disagreement measure introduced in (Gong et al., 2012). We set the Gaussian kernel width parameter σ to the median distance between all source examples. The regularization parameter C of SVM are selected by 5-fold cross validation on a grid. The classification accuracy is given in Table 2. It can be seen that our methods generally work better than DIP and other competitors, which verifies that our methods successfully find the conditional transferable components. Note that the class ratio changes slightly across domains, the main improvement on this dataset and the following WiFi dataset is attributed to the location-scale transform.

5.3. Cross-Domain Indoor WiFi Localization

We finally perform evaluations on the cross-domain indoor WiFi location dataset provided in (Kai Zhang et al., 2013). The WiFi data were collected from the hallway area of an academic building. The hallway area was discretized into a space of 119 grids at which the strength of WiFi signals received from D access points were collected. The task is to predict the location of the device from the D-dimensional WiFi signals, which is usually considered as a regression problem. In our CTC method, we consider Y as a discrete variable when matching the distributions. The training and test data often have different distributions because they are collected at different time periods by different devices.

We compare CTC and CTC-TIP with KMM, surrogate kernels (SuK) (Kai Zhang et al., 2013), TCA, DIP and DIP-CC, and GeTarS. Following the evaluation method in (Kai Zhang et al., 2013), we randomly choose 60% of the examples from the training and test domains and report the average performance of 10 repetitions. The reported accuracy is the percentage of examples on which the predicted location is within 3 meters from the true location. The hyperparamters, including Gaussian kernel width, KRR regularization parameter, and the dimension of W, are tuned on a very small subset of the test domain.

Transfer Across Time Periods In this task, the WiFi data were collected in three different time periods $t1$, $t2$, and $t3$ in the same hallway. We evaluate the methods on three domain adaptation tasks, i.e., t1 → t2, t1 → t3, and t2 → t3. The results are given in the upper part of Table 3. We can see that our methods outperform the IC-type methods like TCA and DIP. Also, our methods are comparable to SuK which is a state-of-the-art method on this dataset.

Transfer Across Devices The signals from different devices may vary from each other due to different signal sensing capabilities. To transfer between different devices, the WiFi data were collected from two different devices at 3 straight-line hallways, resulting in three tasks, i.e., hallway1, hallway2, hallway3. The lower part of Table 3 gives the experimental results. Our methods significantly outperform the competitors, indicating that CTC is very suitable for transferring between devices.

6. Conclusion

We have considered domain adaptation by learning conditional transferable components in the situation where the distribution of the covariate and the conditional distribution of the target given the covariate change across domains. We have shown that, if target causes the covariate, under appropriate assumptions, we are able to find conditional transferable components whose conditional distribution given the target is invariant after proper location-scale transformations, and estimate the target distribution of the target domain. Also, we discussed the relation of our method to the IC-type methods, pointing out that those methods can be considered as a way to achieve our method when the distribution of the target does not change. Finally, we provided theoretical analysis and empirical evaluations to show the effectiveness of our method.

Acknowledgments

The authors thank Kai Zhang for providing the WiFi data. Gong M., Liu T., and Tao D. were supported by Australian Research Council Projects DP-140102164, FT-130101457, and LE-140100061.

References

Baker, C. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–211, 1973.

Baktashmotlagh, M., Harandi, M.T., Lovell, B.C., and Salzmann, M. Unsupervised domain adaptation by domain invariant projection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 769–776, Dec 2013. doi: 10.1109/ICCV.2013.100.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 20*, Cambridge, MA, 2007. MIT Press.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *NIPS 23*, 2010.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, April 1999. ISSN 0895-4798.

Fukumizu, K., Bach, F. R., Jordan, M. I., and Williams, C. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2066–2073. IEEE, 2012.

Gong, B., Grauman, K., and Sha, F. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of The 30th International Conference on Machine Learning*, pp. 222–230, 2013.

Gopalan, R., Li, R., and Chellappa, R. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 999–1006. IEEE, 2011.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL `http://authors.library.caltech.edu/7694`.

Huang, J., Smola, A., Gretton, A., Borgwardt, K., and Schölkopf, B. Correcting sample selection bias by unlabeled data. In *NIPS 19*, pp. 601–608, 2007.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194, 2010.

Jiang, J. *A literature survey on domain adaptation of statistical classifiers*, 2008. URL `http://sifaka.cs.uiuc.edu/jiang4/domain\_adaptation/survey`.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, March 2016.

Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. Transfer joint matching for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1410–1417. IEEE, 2014.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In Blei, David and Bach, Francis (eds.), *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 97–105. JMLR Workshop and Conference Proceedings, 2015. URL `http://jmlr.org/proceedings/papers/v37/long15.pdf`.

Luo, Y., Liu, T., Tao, D., and Xu, C. Decomposition-based transfer distance metric learning for image classification. *IEEE Transactions on Image Processing*, 23(9):3789–3801, Sept 2014. ISSN 1057-7149. doi: 10.1109/TIP.2014.2332398.

Mateo, R., Schölkopf, B., Turner, R., and Peters, J. Causal transfer in machine learning. arXiv:1507.05333, Feb 2016.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning, JMLR: W&CP Vol. 28*, 2013.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22: 1345–1359, 2010.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22:199–120, 2011.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010*, pp. 213–226. Springer, 2010.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *Proc. 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland, 2012.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.

Si, S., Tao, D., and Geng, B. Bregman divergence-based regularization for transfer subspace learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22 (7):929–942, July 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.126.

Si, S., Liu, W., Tao, D., and Chan, K. P. Distribution calibration in riemannian symmetric space. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):921–930, Aug 2011. ISSN 1083-4419. doi: 10.1109/TSMCB.2010.2100042.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer-Verlag, 2007.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12: 2389–2410, 2011.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Yu, Y. and Szepesvári, C. Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 607–614, 2012.

Kai Zhang, Zheng, V., Wang, Q., Kwok, J., Yang, Q., and Marsic, I. Covariate shift in hilbert space: A solution via sorrogate kernels. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 388–395, 2013.

Kun Zhang, Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning, JMLR: W&CP Vol. 28*, 2013.

Kun Zhang, Gong, M., and Schölkopf, B. Multi-source domain adaptation: A causal view. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.